

INDUSTRIAL PRACTICES FINAL REPORT
Online Social Network Message Anonymization
at National Institute of Technology, Gifu College

Proposed for fulfilling the pass requirements of
IF4090 Industrial Practice Course

by :
Gerald Dzakwan / 13514065



INFORMATICS ENGINEERING STUDY PROGRAM
SCHOOL OF ELECTRICAL ENGINEERING & INFORMATICS
BANDUNG INSTITUTE OF TECHNOLOGY
2017

Informatics Engineering Study Program Approval Sheet

Online Social Network Message Anonymization

at National Institute of Technology, Gifu College

by :

Geraldi Dzakwan / 13514065

approved and validated as

Industrial Practice Final Report

Bandung, ____ 2017

ITB Informatics Engineering Study Program Industrial Practice Supervisor

Dr. Eng. Ayu Purwarianti, S.T., M.T.

NIP: 19770127 200801 2 011

Approval Sheet

Online Social Network Message Anonymization

at National Institute of Technology, Gifu College

by :

Geraldi Dzakwan / 13514065

approved and validated as

Industrial Practice Final Report

Motosu, ____ July 2017

Electrical and Computer Engineering Department Professor

Prof. Makoto Yasuda

ID :

Abstract

This final report is for my industrial practice at National Institute of Technology, Gifu College. I did research about Online Social Network Message Anonymization. This type of anonymization is relatively new because it takes raw text (social media message) as the input instead of structured input like commonly found in k-anonymization techniques (e.g. for anonymizing medical record database). The research is done at Artificial Intelligence lab under the supervision of Prof. Makoto Yasuda.

The method used in the research mainly involve machine learning and natural language processing techniques. The first step is to do some preprocessings for the text (e.g. stemming) to clear unuseful information. Next step is to determine the named entities from raw text. I did some experiments using several classifiers (Naive Bayes, Perceptron, etc.) and using state of the art algorithm (Conditional Random Field). I ended up choosing CRF algorithm as it produces better result. Then, for each type of named entity there will be it's own mechanism to determine whether it is a private phrase using steps like executing rule-based approach and calculating co-occurrence metrics with user profile. Next, the detected private named entity will be anonymized in a specific way according to it's type. Finally, there are some post processings to produce the final anonymized message.

The result of the program is good at detecting named entity, with the accuration above 80%. But, the private named entity detection still needs to be improved since there are many other cases/patterns to cover. Often in my program, there are cases when private phrases are not anonymized or vice versa (non-private phrases are anonymized). One of the solution that may works is to apply co-reference resolution to the private named entity detection step. It basically will determine all the words that refer to the same thing and it may improves the private named entity detection performance.

To sum up, I want to give conclusion about the research substance. There are currently not many anonymization research focusing on raw text. Thus, I hope that this research could be extended furthermore in the future in term of improving performance by applying other methods and also in term of covering other domains/areas. I hope that I can also create similar researchs for other domains, such as for raw company or government documents.

Keywords: online social network, text anonymization, social media message, classifier, conditional random field, private named entity, co-occurrence metrics, rule-based approach, phrase generalization, machine learning, natural language processing

Preface

I would like to express my biggest thanks to my internship supervisor, Prof. Makoto Yasuda as he already gave me an opportunity to study in his lab and he also gave me guidances and feedbacks to improve my work results.

I would also like to give my deepest gratitudes to Dr. Eng. Yoshito Ito as the principal of NIT-GC who warmly welcomed me and JASSO which supported me financially to fulfill all my daily expenses while staying in Japan.

My thanks also go to my industrial practice supervisor at ITB, Dr. Eng. Ayu Purwarianti, S.T., M.T. as she had introduced me to this opportunity and she always supported me since the preparation for the internship, including documents preparation for Japan visa and for JASSO scholarship.

I would also not forget to give my thanks to all NIT-GC staffs who helped me out in the administrative processes (Prof. Koji Tajima, Mr. Michael Makoto Martinsen, and Mr. Isao Tomita) and to all my friends in the lab (Hideki Kano, Ryosuke Okachi, etc.) which I couldn't mention all individually. Thank you for taking me around Japan and helping me to find the way out of problems I faced during the internship.

Motosu, 19 July 2017

Geraldi Dzakwan

Chapter I

Introduction

This chapter contains introduction about the industrial practice, including background, scope, and goal.

I.1 Background

Generally, this internship program aims to develop basic research skill in some specific fields that depends on which laboratory the student will choose and what research topic the student will pick. In addition, this internship program is expected to be a bridge between the two countries, Indonesia and Japan, in order to maintain a good relation, especially between Bandung Institute of Technology (ITB) and NIT, Gifu College. This program is also expected to improve the students' communication skill from both schools especially in English.

Regarding the research subject, the student will be given several research topics from each laboratory. Then, the students need to choose their preferred laboratory and their preferred topic or they can also propose their own topic (if approved by the laboratory professor). Afterwards, the student will do research project based on the selected topic. The research project will be conducted for 5 to 6 weeks and it will be supervised by the laboratory professor directly. At the end of the internship program, the student shall submit his internship report to the internship supervisor in NIT, Gifu College.

This section (background) is referred from the Term of Reference (TOR) attachment chapter one (objectives).

I.2 Scope

In the Electrical and Computer Engineering Department of NIT-GC, there are several laboratories with their corresponding research areas. For this term of internship, there are two laboratories which accept ITB students. They are Prof. Tajima's laboratory and Prof. Yasuda's laboratory. Prof. Tajima's laboratory

research areas are Computer Network and IoT. Some research topics which are covered are Online to Offline Computing System, System Development for Linguistic Landscape, and Data Visualizing System. Meanwhile, Prof. Yasuda's laboratory research area is Artificial Intelligence. Some research topics which are covered are Genetic Algorithm, Metaheuristics, and Fuzzy Logic.

My research topic is "Online Social Network Message Anonymization". Its scope is within machine learning and natural language processing. Thus, my research topic belongs to Prof. Yasuda's laboratory (as ML and NLP are subset of AI).

I.3 Goal

The main goal of this research project is to create an anonymization program (API) for social media message which has good accuracy on determining named entities and on determining private phrases. In this context, private phrases all are phrases that may lead to or may identify a person's identity and phrases that are containing user's private information (such as location). Despite making the message anonymized, the message still has to be useful in some ways so it can be used for other purposes such as analytics by third parties. So, there are two parameters that should be satisfied by the result (the anonymized message). The result has to guarantee that it close all user's private information that can lead to his/her identity but at the same time it has to do minimum number of replacement so the result is not far from the original message and the result remains useful.

For further implementation, this anonymization API can be used by messaging applications (such as LINE, WhatsApp, Telegram) to share their message data to third parties. Normally, in today's practice, messaging applications need to strictly encrypt all the messages so the information can't be stolen by third parties. But, actually there are many third parties who want to take benefits from social media messages for doing analytics. The analytics result can be used for supporting some business intelligence processes (e.g. for taking decision in digital marketing). Usually, third parties use public data such as user's tweets to do analytics. But, with this anonymization API they can legally use social media messages (that are already anonymized) for the sake of their analytics. Perhaps this new dataset

(social media messages) can bring improvements to their analytics. Meanwhile, messaging applications can also take benefit by monetizing the messages data (e.g. third parties are charged per request made). So, I think this research topic could lead to something good in the future.

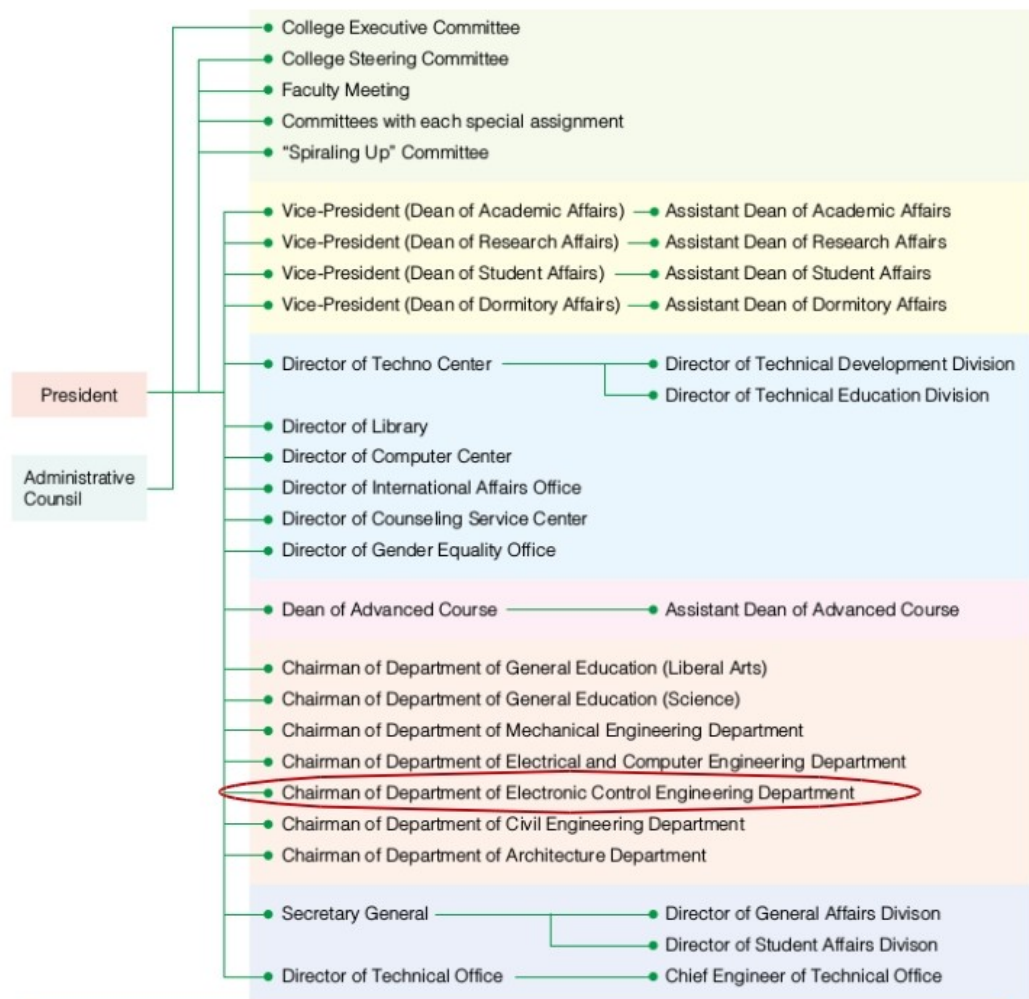
Chapter II

Industrial Practice Work Environment

This chapter contains the details of the workplace (organization) and its work environment, including organization structure of Gifu National College of Technology, project scope, project description, and project schedule.

II.1 Organization Structure

Generally, NIT-GC organization structure is the same as other colleges organization structure. It consists of some administrative committees and seven departments. Below is the illustration of NIT-GC organization structure.



Picture II.1.1 NIT-GC Organization Structure

The lab I was working at (AI Lab, Prof. Makoto Yasuda) is the part of Electrical and Computer Engineering department, highlighted by a red ellipse. Another lab belongs to this department is Computer Network Lab, under Prof. Koji Tajima.

II.2 Project Scope

The internship was held at Artificial Intelligence Laboratory and supervised by Prof. Makoto Yasuda. This laboratory is a part of Electrical and Computer Engineering Department as explained in the previous section. This laboratory is responsible for conducting research around AI areas, such as genetic algorithm, metaheuristics, and fuzzy algorithm.

The research that was conducted during the internship program is related to artificial intelligence, especially natural language processing and machine learning. Hence, I was placed in Prof. Yasuda's laboratory. The research topic is "Online Social Network Message Anonymization". I carried out two projects, which are training named entity recognizer and creating message anonymization program (API). The description for each project will be explained in the next section.

This section (project scope) is referred from the Term of Reference (TOR) attachment chapter two (project scopes).

II.3 Project Description

As mentioned before in the previous section, there are two main projects in the research and below is the description of both projects, explained step by step.

a. Training Named Entity Recognizer

In this project, the student should be able to make a program for predicting named entities from an input word/phrase. The program will use a trained model that will be built using Python Scikit-Learn module (a popular machine learning module in Python). The student will do several

experiments using different classifier or algorithm and different data train sample. The student will then decide which model is to be used in the message anonymization API based on the accuracy testing. For the dataset which will be used in training, the student will use Groningen Meaning Bank corpus as so far it is the biggest, most complete, and most suitable English corpus the student has found.

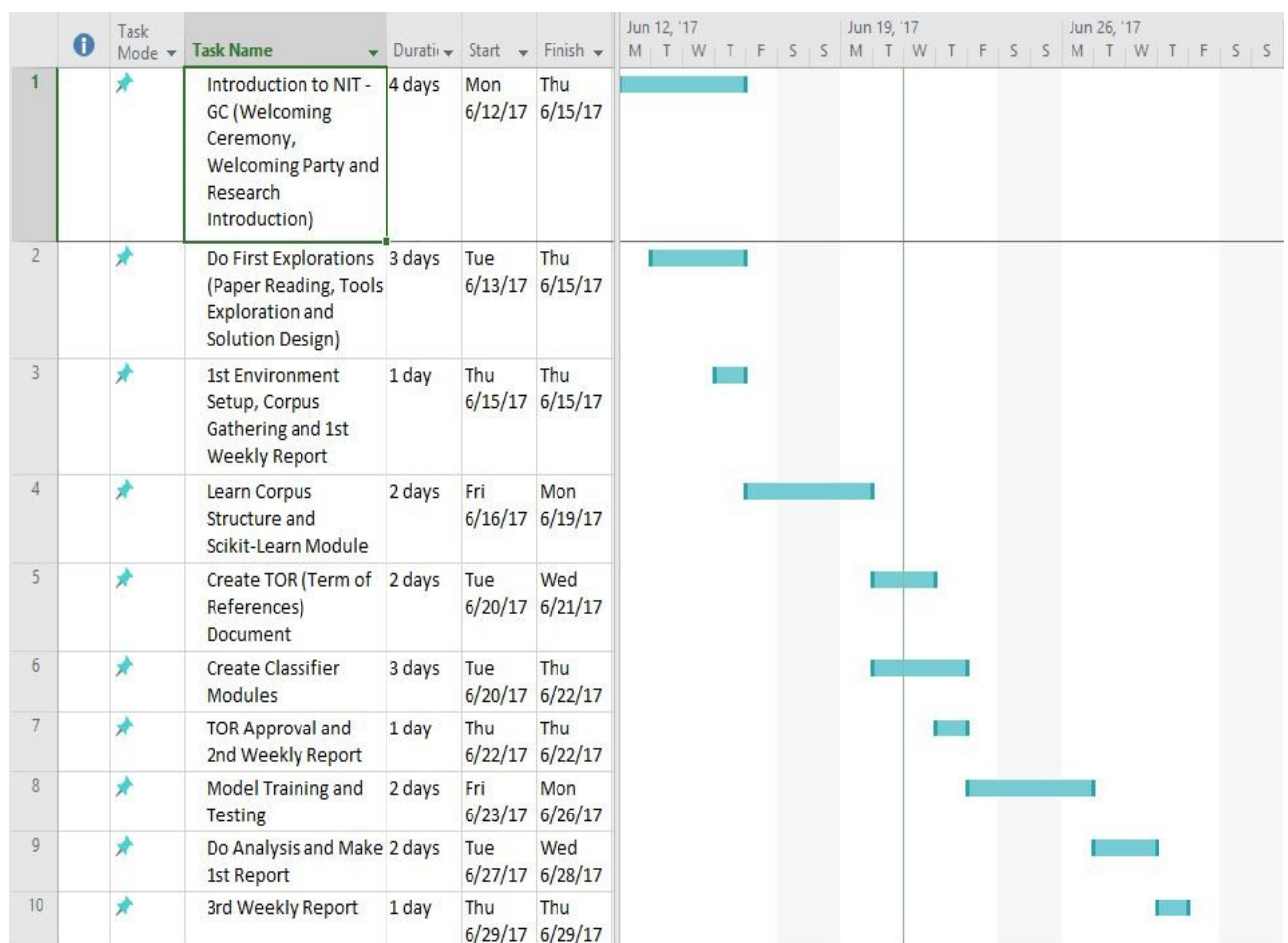
b. Creating Message Anonymization Program (API)

This project is the next step after the first project. This project will use the model which has been created in the first project. The program will probably be in the form of web API, built using either Flask or Django framework. This project will receive an input message (a private message, consisting of one or more sentences, using JSON format) from a client via HTTP POST request. Then, it will do some pre-processing procedures (e.g. sentence tokenizing, POS tagging, NP-chunking and punctuation removal). The result of preprocessing will be delivered to the classifier. The classifier then will return the sentences back to the program with identified named entities. Afterwards, the program will do some post-processing procedures (e.g. correction, co-reference resolution, and sentence restructuring). Last but most important procedure is to apply the anonymization. There are two ways of doing the anonymization, the first one is to replace the named entities with their class or to replace it with another phrase which belongs to the same class so it becomes more natural. The student will try to explore both options. The restructured sentences which have been anonymized then be sent back to the client via JSON message response.

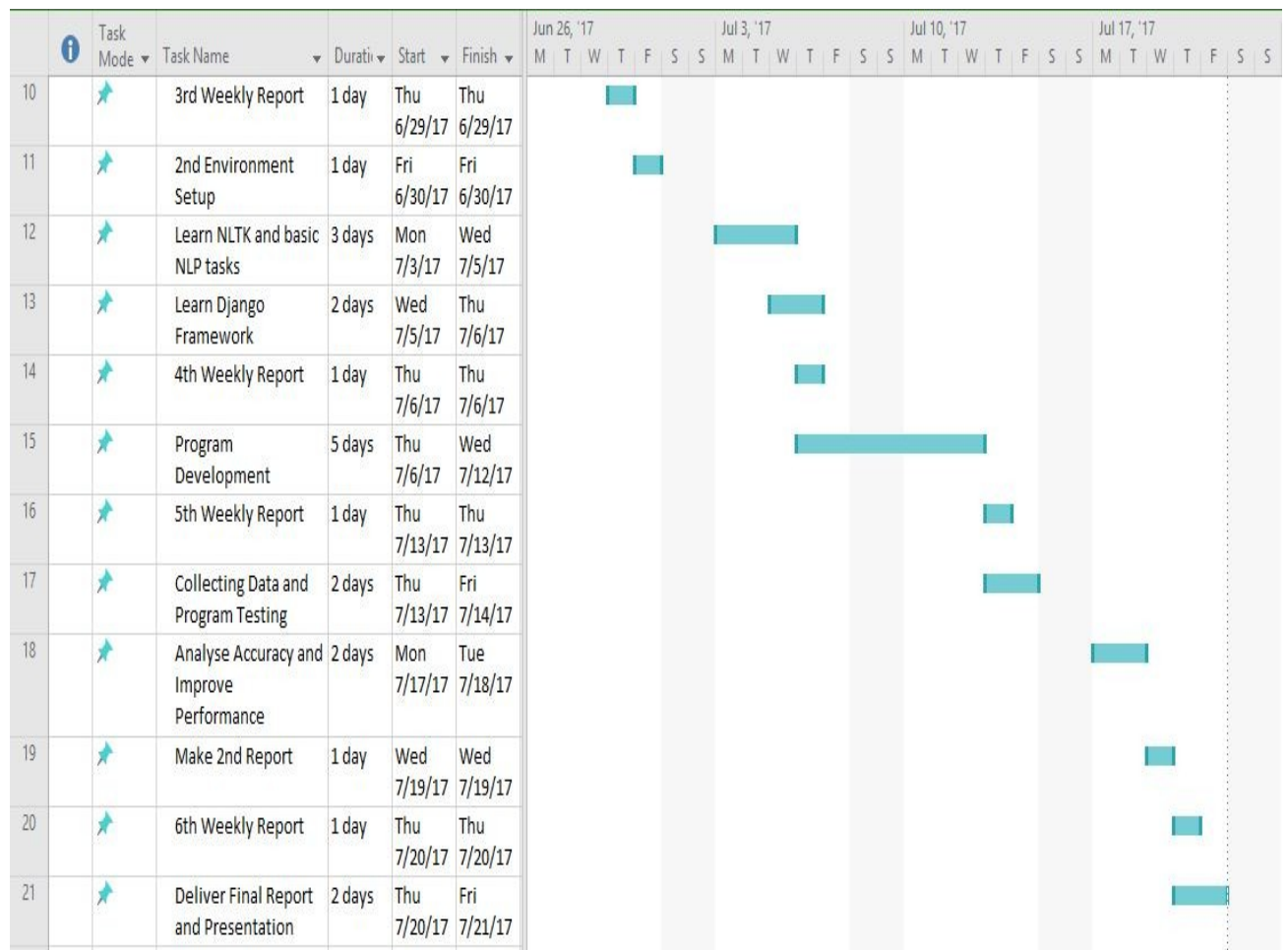
This section (project description) is referred from the Term of Reference (TOR) attachment chapter two (project scopes).

II.4 Project Schedule

Generally, the student must come to the laboratory every work day (Monday-Friday), from 9.00 a.m. until 16.10 p.m. to work on the research project. Despite that, other activities might take place during research time, such as attending welcoming ceremony and welcoming party, weekly talk café (interacting with NIT-GC students) and giving presentation in some classes. The detailed schedule can be seen below. There are two pictures, the first one shows Gantt chart for the first project and the second one shows Gantt chart for the second project.



Picture II.4.1 Project 1 Gantt Chart



Picture II.4.2 Project 2 Gannt Chart

This section (schedule) is referred from the Term of Reference (TOR) attachment chapter eight (schedule).

Chapter III

Online Social Network Message Anonymization Implementation

This chapter contains the explanation of how the project is done, including problem description, solution design (step-by-step), implementation, difficulties faced, result, and analysis.

III.1 Problem Description

There are two main backgrounds/reasons why I choosed this research topic. The first one is because I want to get some hands on about machine learning and natural language processing. The second one is because there are currently not many anonymization research focusing on raw text. This type of anonymization is different because it takes raw text (social media message) as the input instead of structured input like commonly found in k-anonymization techniques (e.g. for anonymizing medical record database).

The problem description is best explained by stating the research goal itself. The goal is to create an anonymization program (API) for social media message which has good accuracy on determining named entities and on determining private phrases. In this context, private phrases all are phrases that may lead to or may identify a person's identity and phrases that are containing user's private information (such as location). Despite making the message anonymized, the message still has to be useful in some ways so it can be used for other purposes such as analytics by third parties. So, there are two parameters that should be satisfied by the result (the anonymized message). The result has to guarentee that it close all user's private information that can lead to his/her identity but at the same time it has to do minimum number of replacement so the result is not far from the original message and the result remains useful.

Regarding the research project, there are some constraints.

The followings are some constraints that are applied throughout the first project about training named entity recognizer:

1. The programming language used is Python version 2.7.10.
2. The English corpus dataset used is Groningen Meaning Bank (GMB) corpus version 2.2.0.
3. The main machine learning tools used for training the model is Python Scikit-Learn version 0.17.
4. The classes of named identity that can be identified are limited to: person name, location/geographical entity, organization, geopolitical entity, and time indicator.
5. Accuracy testing will be done by split test, which is separating train dataset and test dataset from the corpus.

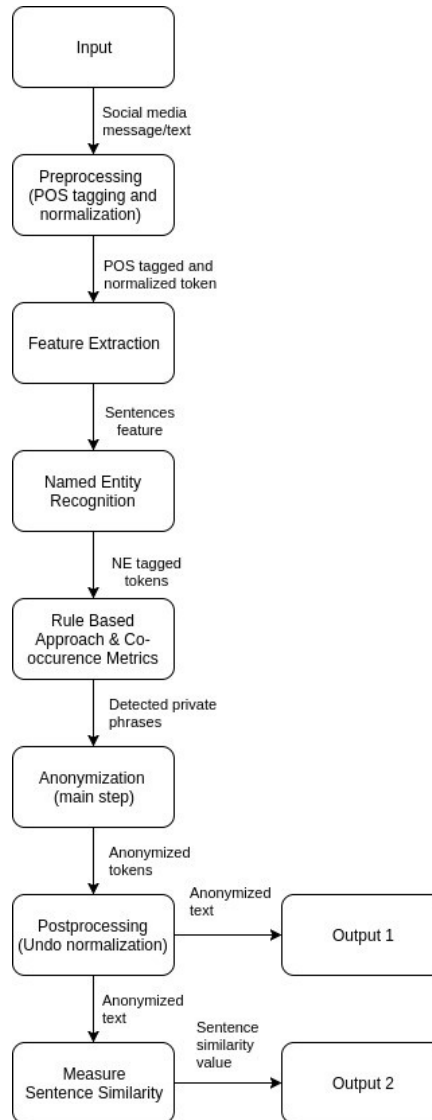
The followings are some constraints that are applied throughout the second project about creating message anonymization program:

1. The programming language used is Python version 2.7.10.
2. The main tools for doing natural language processing techniques are Python NLTK (Natural Language Toolkit) version 3.2.4.
3. The framework used for creating web API is either Flask or Django.
4. The program will only receive private message (a message that has to be anonymized) as the input.
5. Testing will be done by data from GMB corpus and other data that are collected manually by the student.

This section (project description) is referred from the Term of Reference (TOR) attachment chapter four (project constraints).

III.2 Solution Design and Implementation

We will first take a look to the big picture of anonymization solution design. In a nutshell, it looks like this:



Picture III.2.1 Solution design scheme

The first step to be done is to do some preprocessings. There are four main things inside the preprocessings, which are tokenization, pos tagging, stemming, and lemmatization. Tokenization is an activity to split a sentence into a sequence of tokens, which roughly correspond to "words". Meanwhile, POS (Part-Of-Speech) Tagging is an activity to read text and assigns parts of speech to each token, such

as noun, verb, adjective, etc. In this project, I use Penn Treebank type of pos tag. The pos tags list can be seen here :

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html .

The tokenization and pos tagging procedure is done by using `nltk.word_tokenize` and `nltk.pos_tag` library. This pos tag will be used later as a feature for other processes, such as lemmatization and named entity recognition.

Next process is stemming and lemmatization. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form. Stemmers use an algorithmic approach of removing prefixes and suffixes. The result might not be an actual dictionary word. For example, the word “fishing”, “fished”, and “fisher” will be stemmed to the root word “fish”. Meanwhile, lemmatization is a kind of similar process. But, lemmatizers always returns a dictionary word as it uses a corpus and it needs extra info about the part of speech they are processing (e.g. “calling” can be either a verb or a noun). Lemmatization even can turn past tense verb to present tense verb (e.g. went to go) but it is way slower than stemmers. The stemming is done by using `SnowballStemmer` library and lemmatization is done by using the `WordNetLemmatizer` library. These two processes are needed to normalize the text and clear unnecessary informations.

Next, I will explain about the named entity classifier, a very important part of this project. The main corpus dataset that I used is Groningen Meaning Bank classifier version 2.2.0 which could be found here (<http://gmb.let.rug.nl/data.php>). In the classification, I define eight classes/entities, which are :

1. per (person’s name). Example : ‘John Smith’ , ‘Alice’
2. geo (location). Example : ‘Gifu’, ‘Motosu’
3. org (organization). Example : ‘Google’, ‘Bandung Institute of Technology’
4. tim (time indicator). Example : ‘at 2 pm’, ‘this Sunday’
5. gpe (geopolitical entity). Example : ‘Japan’, ‘Japanese’
6. eve (event). Example : ‘All England Championship’, ‘Thailand Open’
7. art (artifact). Example : ‘house’, ‘dollar’

8. nat (natural phenomenon). Example : 'disease', 'storm'

But, only the first six entities would be processed further to detect private phrases. That is because the last two entities are arguably not private informations, thus they don't have to be anonymized. Before moving further, I would want to explain about a concept called IOB tag (a popular concept in extracting information from text) which described as below.

1. B-named_entity_tag means that it is the first word of named entity.
2. I- named_entity_tag means that it is the word inside a chunk/entity.
3. O means that the word is outside of chunk or not a part of named entity.

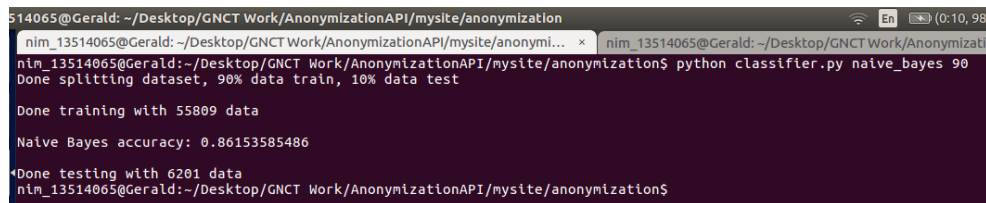
For example, if there is a sentence 'John Smith is a student'. Then, the word John would be tagged as B-person, the word Smith would be tagged as I-person and the rest of the words are tagged as O.

For the classifier, my first try is to use Naive Bayes and Perceptron classifier. There are some features that I extracted from the words in each sentence in the dataset to be learned by the classifier. Some important features used are:

1. The word itself
2. The word's lemma
3. The word position in the sentence
4. The word pos tag
5. If the word contains dash
6. If the word contains digit
7. If the word is capitalized
8. If all the letter in the word is capital
9. The previous word IOB tag (most likely there is transition from B-named_entity_tag to I-named_entity_tag) if the word position is not zero (not the first word).

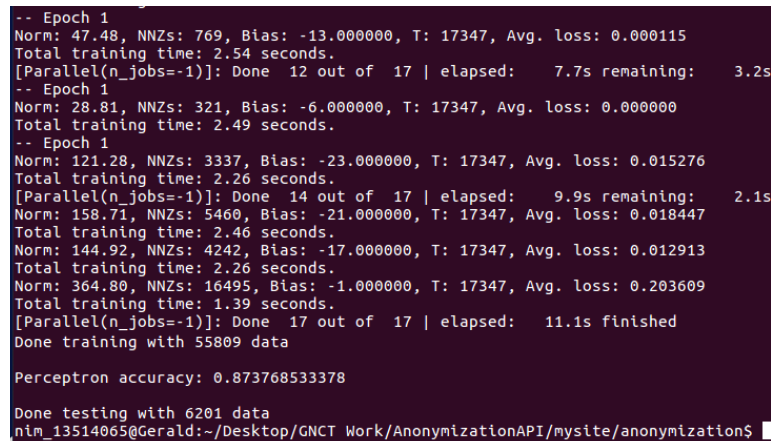
Accuracy testing is done by splitting the data train and data test. Data train consists of 90% of whole dataset while data test consists of 10% of whole dataset. Below is the result for each classifier.

Naive Bayes accuracy:

A terminal window with a dark background and light text. The prompt is 'nim_13514065@Gerald:~/Desktop/GNCT Work/AnonymizationAPI/mysite/anonymization\$'. The command entered is 'python classifier.py naive_bayes 90'. The output shows 'Done splitting dataset, 90% data train, 10% data test', 'Done training with 55809 data', and 'Naive Bayes accuracy: 0.86153585486'. The prompt is then 'nim_13514065@Gerald:~/Desktop/GNCT Work/AnonymizationAPI/mysite/anonymization\$'.

Picture III.2.2 Naive Bayes classifier accuracy

Perceptron accuracy:

A terminal window with a dark background and light text. The prompt is 'nim_13514065@Gerald:~/Desktop/GNCT Work/AnonymizationAPI/mysite/anonymization\$'. The command entered is 'python classifier.py perceptron 90'. The output shows multiple epochs of training with metrics like Norm, NNZs, Bias, T, and Avg. loss. The final output is 'Perceptron accuracy: 0.873768533378'. The prompt is then 'nim_13514065@Gerald:~/Desktop/GNCT Work/AnonymizationAPI/mysite/anonymization\$'.

Picture III.2.3 Perceptron classifier accuracy

We might see that the accuracy is pretty high (above 80%). But, actually it isn't as there are only 10-20% of words in the dataset that have named entity/class. Most of the words (>80%) in dataset are not named entity or tagged as O (don't belong to one of the eight classed defined). If we predict that all words have O tag (again, O stands for outside any entity), we will achieve at least 80% accuracy. Thus, this named entity classifier is quite poor (not really good).

So, I try to find another method/algorithm to do the named entity recognition. After exploring some alternatives, I ended up choosing an algorithm called Conditional Random Field. Conditional random fields (CRFs) are a class of statistical modeling method often applied in pattern recognition and machine learning and used for structured prediction. CRFs fall into the sequence modeling family. Whereas a discrete classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account; e.g., the

linear chain CRF (which will be used) predicts sequences of labels for sequences of input samples.

There are some features that I extracted from the words in each sentence in the dataset to be learned by the CRF classifier. Some important features used are:

1. The word itself
2. The word's lemma
3. The word position in the sentence
4. The word pos tag
5. The last two letters
6. The last three letters
7. If the word contains digit
8. If the word contains dash
9. If the word is capitalized
10. If all the letter in the word is capital
11. The IOB tag

If the word is not the first word in the sentence, then there will be added features which are some of the above features for the previous word (the previous word's lemma, the previous word pos tag, the previous word IOB tag, if the previous word is capitalized, etc.). If the word is the first word then there will be one added feature that states it is the beginning of the sentence (BOS).

If the word is not the last word in the sentence, then there will be added features which are some of the above features for the next word (the next word's lemma, the next word pos tag, the next word IOB tag, if the next word is capitalized, etc.). If the word is the last word then there will be one added feature that states it is the end of the sentence (EOS).

The advantage of using CRF is that it uses sentence pattern as its feature. This algorithm computes the probability of transitions between classes. For example, in my case, the algorithm learns that it is more likely to transit from B-ORG to I-ORG than from B-ORG to any other classes. This is perhaps because in the

dataset I used, organization entities usually consist more than a word. This is the detail of most likely and unlikely transitions.

```

Likely transitions:
I-org -> I-org 5.911407
B-per -> I-per 5.703036
B-org -> I-org 5.658637
I-per -> I-per 5.283008
I-nat -> I-nat 4.841115
O -> O 4.113861
O -> B-per 2.411824
B-gpe -> B-org 2.331100
B-org -> B-art 1.959518
B-gpe -> O 1.859497

Unlikely transitions:
O -> I-art -3.812120
B-geo -> I-org -3.890835
I-org -> B-org -3.946717
I-org -> I-per -4.023397
B-gpe -> I-org -4.034168
B-org -> B-org -4.140429
B-tim -> B-tim -4.168018
B-org -> I-per -4.359579
I-per -> B-per -5.083065
nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization$

```

Picture III.2.4 CRF transitions

As the result, CRF model has higher accuracy than the two previous models. Below are the accuracy and confusion matrix of the CRF model. The first picture depicts testing including words with O label while the second picture depicts testing ignoring words with O label.

```

nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization$ python3
/usr/local/lib/python2.7/dist-packages/sklearn/cross_validation.py:42: DeprecationWarning:
  "This module will be removed in 0.20.", DeprecationWarning)
/usr/local/lib/python2.7/dist-packages/sklearn/grid_search.py:77: DeprecationWarning:
  "This module will be removed in 0.20.", DeprecationWarning)

Accuracy:
0.971549402036

Confusion matrix:
precision    recall  f1-score   support

   0           0.992    0.994    0.993    115381
  B-art        0.238    0.106    0.147         47
  I-art        0.143    0.018    0.032         55
 B-eve        0.625    0.424    0.505         59
 I-eve        0.433    0.317    0.366         41
 B-geo        0.859    0.910    0.884        4880
 I-geo        0.825    0.823    0.824         971
 B-gpe        0.967    0.946    0.956        2146
 I-gpe        0.889    0.421    0.571         19
 B-nat        0.556    0.263    0.357         19
 I-nat        0.667    0.286    0.400          7
 B-org        0.790    0.727    0.757        2548
 I-org        0.820    0.796    0.807        2032
 B-per        0.858    0.823    0.841        2288
 I-per        0.852    0.903    0.876        2331
 B-tim        0.929    0.885    0.906        2695
 I-tim        0.839    0.786    0.811         900

avg / total           0.971    0.972    0.972    136419

```

Picture III.2.5 CRF model accuracy including 'O' label

```

Accuracy:
0.853137168992

Confusion matrix:
precision    recall  f1-score   support

 B-art      0.238    0.106    0.147     47
 I-art      0.143    0.018    0.032     55
 B-eve      0.625    0.424    0.505     59
 I-eve      0.433    0.317    0.366     41
 B-geo      0.859    0.910    0.884    4880
 I-geo      0.825    0.823    0.824     971
 B-gpe      0.967    0.946    0.956    2146
 I-gpe      0.889    0.421    0.571      19
 B-nat      0.556    0.263    0.357      19
 I-nat      0.667    0.286    0.400       7
 B-org      0.790    0.727    0.757    2548
 I-org      0.820    0.796    0.807    2032
 B-per      0.858    0.823    0.841    2288
 I-per      0.852    0.903    0.876    2331
 B-tim      0.929    0.885    0.906    2695
 I-tim      0.839    0.786    0.811     900

avg / total      0.859    0.850    0.853    21038

nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization

```

Picture III.2.6 CRF model accuracy ignoring ‘O’ label

We can see that there is a quite significant accuracy difference between both of them. The more valid one is the second accuracy because it ignores the testing for all ‘O’ label that are populating the dataset which I considered as ‘trash’. Hence, I could say my CRF named entity recognition accuracy is above 80% in GMB dataset only.

I use the CRF classifier to predict the named entity of each token. There are three possibilities, which are B tag, I tag, or O tag. For each consecutive B-tag and I tag, the program will combine all the chunks/tokens to become one whole chunk/token. For example, the sentence ‘John Smith is a student’ will be tagged as below.

John, B-per – Smith, I-per – is, O – a, O – student, O.

Then, the program will combine all consecutive B-tag and I-tag and it will look like this afterwards.

John Smith, per – is, O – a, O – student, O.

As stated before, the goal of the project is the anonymized message should guarantee that it close all informations that can identify a person. But, in the same time it has to do minimum number of replacement so the result is not far from the

original message and the result remains useful. Thus, there must be a mechanism which make sures only private word/phrase that are replaced. So, I proposed two steps to achieve those goals. The first step is to do rule based approach to determine the private phrases. The second step is to calculate co-occurrence metrics between detected nouns to user profile.

The first step is the rule based approach. The rule for each entity has a little bit difference but they have a based common framework. The steps for rule based approach are as below.

1. Extracting candidate phrases

A candidate phrase p is defined as a phrase that begins with the word “I” or “My” and ends with a locational entity, personal entity, organizational entity, geopolitical entity, or event entity. p is also checked whether it contains dot (‘.’) as full stop punctuation. If it does, then it is excluded from candidate phrases. This is needed to avoid creating candidate phrases from more than one sentence. The candidate phrase p extracted will be processed further to the next steps.

2. Checking negative phrases

p is checked if it has negative meaning or includes negative words (like not, no, ain’t, and never). If p has negative meaning then it will not be processed further as it doesn’t contain private information. Otherwise, if p doesn’t have negative meaning, then it will be processed further to the next steps.

3. Checking non-private verbs/nouns

p is checked if it contains at least one non-private locational verbs/nouns, non-private personal verbs/nouns, non-private organizational verbs/nouns, non-private geopolitical verbs/nouns, or non-private event verbs/nouns. These verbs/nouns are stored in the MySQL database so it can be extended easily. If p contains non-private verbs/nouns then it will not be processed further as it doesn’t contain private information. Otherwise, if p doesn’t

contain non-private verbs/nouns, then it will be processed further to the next steps.

4. Checking private verbs/nouns

p is checked if it contains at least one private locational verbs/nouns, private personal verbs/nouns, private organizational verbs/nouns, private geopolitical verbs/nouns, or private event verbs/nouns. These verbs/nouns are stored in the MySQL database so it can be extended easily. If p contains private verbs/nouns then it will be treated as private phrases. Otherwise, if p doesn't contain private verbs/nouns, then it will not be treated as private phrases as it doesn't contain private information.

Below are sample verbs/nouns for locational entities.

```
mysql> SELECT * from private_locational_verbs LIMIT 7;
+-----+
| word |
+-----+
| live |
| stay |
| fly  |
| go   |
| hometown |
| address |
| meet  |
+-----+
7 rows in set (0,00 sec)

mysql> SELECT * from non_private_locational_verbs LIMIT 7;
+-----+
| word |
+-----+
| like |
| love |
| want |
| hear |
| feel |
| think |
| ask  |
+-----+
7 rows in set (0,00 sec)
```

Picture III.2.7 Sample verbs/nouns for locational entities

The temporal/time entity is approached in a rather different way. I found that most temporal entity comes in a fix pattern. So, I create a regex-based temporal entity recognition. Here are some examples of basic patterns that I used:

1. day_regex = day_preposition (on | this | next | last) + days (Monday .. Sunday) + day_details* [morning | afternoon | evening | night]. Some phrases that match this regex: this Sunday, next Saturday, on Monday morning.

2. `hour_regex_1 = hour_preposition (at) + numbers [0..23] + conjunction* [., | :] + numbers* [0..59] + hour_desc* [am | pm]`. Some phrases that match this regex: at 2.45 pm, at 2 pm.
3. `hour_regex_2 = hour_preposition (at) + hour [one .. twelve] + minute [thirty | fourty five | fifteen | etc.] + hour_desc* [am | pm]`. Some phrases that match this regex: at ten thirty pm, at one pm

These regexes can be combined to create another regex such as `day_and_hour = day_regex + hour_regex_1 | hour_regex_2` to cover larger phrases like ‘next Monday evening at 3.40 pm’.

The goal of the second step is to check if there are other entities that are not detected as named entity from the first step and may identify a person (or may be related to user profile). Because entities are always pos tagged as nouns, I do noun phrase (NP) chunking to the sentences to get all nouns and then exclude all the nouns that have been identified as named entity before.

The method used is to calculate the co-occurrence metrics to determine how close is the connection between all the nouns detected and all aspects of the user profile. I simulate a user profile containing some common informations using a MySQL database and it looks like this:

```
mysql> select * from user_profile;
```

education	work	email_address	full_name	hometown_city	current_city
Bandung Institute of Technology	Qontak	geraldi.dzakwan@gmail.com	Gerald Dzakwan	Jakarta	Gifu
University of Indonesia	Dattabot	alif.karnadi@gmail.com	Alif Karnadi	Jakarta	Jakarta

2 rows in set (0.00 sec)

Picture III.2.8 User profile database

To calculate the co-occurrence metrics, I used Google Custom Search API. It enables me to retrieve the number of web pages that contains at least one of the words or both of the words. The co-occurrence metrics are calculated using the formula below.

$$Co(X, Y) = \frac{Fr(X \cap Y)}{Fr(X \cup Y)}$$

Picture III.2.7 Co-occurrence metrics formula

It is a division result between the number of pages that contains both of the words and the number of pages that contains at least one of the word. A high result means that the two words have a closed connection and vice versa (low result means that the two words have a loose connection or no connection at all). For example, the co-occurrence metrics value of Tokyo and Japan is 0,078 and the co-occurrence metrics value of Tokyo and Korea is 0,047. It means that Tokyo and Japan have a closer connection than Tokyo and Korea as we would expect.

After doing several experiments, I decided to put threshold between 0,01 – 0,03. It means that:

1. If the result exceeds 0,03 , it is categorized as private word.
2. If the result is below 0,01 , it is categorized as non-private word.
3. If it is between 0,01 and 0,03 , it is categorized as a private word with low confidence level.

There are some reasons why I can't train a model or doing some statistical approaches to determine the threshold. It is because Google Custom Search API limits the number of request that can be made per day up to 100 request for free users. Meanwhile, for processing a text it may takes about 6 – 30 requests as there are 6 aspects of user profile and 5 nouns in the text (average). Hence, I can only use the request for the sake of developing and debugging the program in this case. I may look for alternatives dataset/API other than Google in the future to use for further improvements on the threshold and co-occurrence metrics calculation and to make my program more scalable.

After detecting private phrases, anonymizations are done these ways:

1. For personal entity, it would be replaced by another person name with the same gender (woman with woman, man with man). The gender prediction task is done using genderize.io API which is freely available. For the replacement, I create a database containing person's name. The replacement would be taken from the person database randomly.

Single Usage

An example of genderizing a single name could look like this.

```
GET https://api.genderize.io/?name=peter
```

This would render a JSON response like the following. The count represents the number of data entries examined in order to calculate the response.

```
{"name":"peter","gender":"male","probability":"0.99","count":796}
```

Picture III.2.9 Genderize.io usage

2. For locational entity, it would be replaced by another entity that belongs to the same general entity if possible (if there is another entity that belongs to the same general entity). If not possible, it will be replaced by a more general entity up by one level. The level are as below.

- a. Continent
- b. Country
- c. Subdivision 1
- d. Subdivision 2
- e. City

These levels are based on a geographical dataset that can be taken from <https://dev.maxmind.com/geoip/geoip2/geolite2/>. Subdivision 1 and subdivision 2 are similar to a state in USA or to a prefecture in Japan. For example, if there is a private locational word ‘Motosu’ (which is a city), then it would be replaced by ‘Gero’ (which is a city that belongs to same subdivision Gifu) or by ‘a city in Gifu’ (which is a subdivision 1 in the dataset). Below are a sample subset from the dataset.

```
mysql> mysql> SELECT * FROM location WHERE subdivision_1_name = 'Gifu' LIMIT 5;
```

continent_name	country_name	subdivision_1_name	subdivision_2_name	city_name
Asia	Japan	Gifu		Gero
Asia	Japan	Gifu		Ueda
Asia	Japan	Gifu		Toki
Asia	Japan	Gifu		Tarui
Asia	Japan	Gifu		Takayama

```
5 rows in set (0,01 sec)
```

Picture III.2.10 Sample location dataset

3. For organizational entity, it would be ideally replaced by the type of organization/company like ‘tech company’, ‘educational organization’,

etc. But, I haven't figured out how to do this. So, currently, it would be simply replaced by:

1. 'a college' if the organizational entity contains the word 'university', 'institution', 'institute', or 'college' or the phrase contains the word 'studi' or 'study'.
 2. 'an organization' if the organizational entity contains the word 'foundation', 'organization', etc or the phrase contains the word 'member', 'belong', etc.
 3. 'a company' if the organizational entity doesn't match the rules above or the phrase contains the word 'work', 'job', etc.
-
4. For temporal/time entity, it would be replaced by more general temporal phrase based on the specific hour, day, or date. Some examples of replacement are as below.
 1. Exact hour (e.g. 2.30 pm). This will be replaced by one of this expression : morning, afternoon, evening, night.
 2. Exact day (e.g. Monday). This will be replaced by a general expression like 'some day this week'.
 3. Exact date (e.g. 12 January). This will be deleting the date (e.g. in some day in January).

After doing the anonymization, the sentences are still wrapped up as tokens which have been normalized (stemmed and lemmatized). So, things to do are to undo the normalization and rebuild the sentence to return the result as a text. Ideally, there should be some co-reference resolution mechanism here but I haven't got time to accomplish. More on why should there be a co-reference mechanism will be explained in the next section.

As stated before, the anonymization has to do minimum number of replacement so the result is not far from the original message and the result remains useful. So, the last step would be measuring sentence similarity to determine how good is the result. In this project, I used WordNet to compute the sentence similarity using

properties like Synset and POS tag. The algorithm itself is taken from the paper <https://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf>.

III.3 Result and Analysis

In this section, I will give two input examples and explain how my program works to generate the output. The first sentence is ‘My hometown is Jakarta. My favorite food is fried rice. I’ve studied at Bandung Institute of Technology for three years majoring in computer science.’

The first step is to apply the pos tagging, stemming, and lemmatization. We can see that all the words are converted to it’s basic/root word/lemma to make the further processing easier. Some examples of converted word are:

1. ‘is’ becomes ‘be’
2. ‘studied’ becomes ‘studi’
3. ‘majoring’ becomes ‘major’

The result is as below.

```
nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization$ python main.py load "save_model_crf_gmb_dua_kali.pkl" "My hometown is Jakarta. My favorite food is fried rice. I've studied at Bandung Institute of Technology for three years majoring in computer science." "Gerald Dzakwan"
POS Tagging, Stemming, and Lemmatization:
[('my', 'PRPS'), (u'hometown', 'NN'), (u'be', 'VBZ'), ('Jakarta', 'NNP'), ('.', '.'), ('my', 'PRPS'), (u'favorit', 'JJ'), (u'food', 'NN'), (u'be', 'VBZ'), (u'fri', 'VBN'), (u'rice', 'NN'), ('.', '.'), ('i', 'PRP'), (u've', 'VBP'), (u'studi', 'VBN'), ('at', 'IN'), ('Bandung', 'NNP'), ('Institute', 'NNP'), ('of', 'IN'), ('technology', 'NNP'), (u'for', 'IN'), (u'three', 'CD'), (u'year', 'NNS'), (u'major', 'VBG'), ('in', 'IN'), (u'comput', 'NN'), (u'scienc', 'NN'), ('.', '.')]

```

Picture III.3.1 Example 1 step 1

The purpose of this normalization is so that I can compare these words to the private and non-private verbs/nouns in the database. Because, all the verbs/nouns in the database are in their root form and I can’t compare the words in the sentence directly without being normalized first.

The second step is to apply the named entity recognition and combine all the named entity chunks to one whole chunk. In this example, ‘Bandung Institute of Technology’ become one big chunks of organization. The result is as below.

```
Named entity recognition and combining chunks:
['O', 'O', 'O', 'B-geo', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-org', 'I-org', 'I-org', 'I-org', 'O', 'B-tim', 'O', 'O', 'O', 'O', 'O', 'O']

[['my', 'O'], [u'hometown', 'O'], [u'be', 'O'], ['Jakarta', 'geo'], [',', 'O'], ['my', 'O'], [u'favorit', 'O'], [u'food', 'O'], [u'be', 'O'], [u'fri', 'O'], [u'rice', 'O'], [',', 'O'], ['i', 'O'], [u've', 'O'], [u'studi', 'O'], [u'at', 'O'], ['Bandung Institute of Technology', 'org'], [u'for', 'O'], [u'three', 'tim'], [u'year', 'O'], [u'major', 'O'], [u'in', 'O'], [u'comput', 'O'], [u'scienc', 'O'], [',', 'O']]
```

Picture III.3.2 Example 1 step 2

The third step is to detect the private locational candidate phrases and do anonymization with another subdivision/state that belongs to the same country. In this example, the private phrases are ‘my hometown be Jakarta’ because it starts with the word my, ends with locational entity, doesn’t have non-private locational verb, and contains private locational verb ‘hometown’. Furthermore, Jakarta province belongs to Indonesia and thus be replaced/anonymized by another random province, which in this case is Central Java. The result is as below.

```
Private locational candidate phrases:
[['my', u'hometown', u'be', 'Jakarta']]

Anonymize private locational phrases:
[['my', 'O'], [u'hometown', 'O'], [u'be', 'O'], ['Central Java', 'geo'], [',', 'O'], ['my', 'O'], [u'favorit', 'O'], [u'food', 'O'], [u'be', 'O'], [u'fri', 'O'], [u'rice', 'O'], [',', 'O'], ['i', 'O'], [u've', 'O'], [u'studi', 'O'], [u'at', 'O'], ['Bandung Institute of Technology', 'org'], [u'for', 'O'], [u'three', 'tim'], [u'year', 'O'], [u'major', 'O'], [u'in', 'O'], [u'comput', 'O'], [u'scienc', 'O'], [',', 'O']]
```

Picture III.3.3 Example 1 step 3

The fourth step is to detect the private organizational candidate phrases and do the anonymization. In this example, the private phrases are ‘i’ve studi at Bandung Institute of Technology’ because it starts with the word I, ends with organizational entity, doesn’t have non-private organizational verb, and contains private organizational verb ‘studi’. Furthermore, since Bandung Institute of Technology contains the word ‘institute’ and the phrase contains the word ‘studi’, then it is replaced/anonymized by the word ‘college’. The result is as below.

```
Private organizational candidate phrases:
[['i', u've', u'studi', 'at', 'Bandung Institute of Technology']]

Anonymize private organizational phrases:
[['my', 'O'], [u'hometown', 'O'], [u'be', 'O'], ['Central Java', 'geo'], [',', 'O'], ['my', 'O'], [u'favorit', 'O'], [u'food', 'O'], [u'be', 'O'], [u'fri', 'O'], [u'rice', 'O'], [',', 'O'], ['i', 'O'], [u've', 'O'], [u'studi', 'O'], [u'at', 'O'], ['college', 'org'], [u'for', 'O'], [u'three', 'tim'], [u'year', 'O'], [u'major', 'O'], [u'in', 'O'], [u'comput', 'O'], [u'scienc', 'O'], [',', 'O']]
```

Picture III.3.4 Example 1 step 4

The fifth step is to convert list of normalized token back into a text. We have to also convert back all the words that had been converted to their lemma back to the original words. The last step is to compute sentence similarity to calculate the information loss as an important parameter. The result is 0,83 which means the two sentences have high similarity and low information loss. The result is as below.

```
Final anonymized sentence:
My hometown is Central Java . My favorite food is fried rice . I 've studied at college for three years majoring in computer science .

Similarity measure value:
0.833258928571
nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization$
```

Picture III.3.5 Example 1 step 5

The second sentence is ‘I will meet my sister, Alice, at 3 pm maybe around Motosu.’

The first step is to apply the pos tagging, stemming, and lemmatization. We can see that all the words are converted to it’s basic root word/lemma to make the further processing easier. The result is as below.

```
nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization$ python main.py load "save_model_crf_gmb_dua_kali.pkl" "I will meet my sister, Alice,
at 3 pm maybe around Motosu." "Gerald Dzakwan"

POS Tagging, Stemming, and Lemmatization:
[('I', 'PRP'), ('will', 'MD'), ('meet', 'VB'), ('my', 'PRPS'), ('sister', 'NN'), (',', ','), ('Alice', 'NNP'), (',', ','), ('at', 'IN'), ('3', 'CD'), ('pm', 'NN'), ('maybe', 'RB'), ('around', 'IN'), ('Motosu', 'NNP'), ('.', '.')]

```

Picture III.3.6 Example 2 step 1

The second step is to apply the named entity recognition and combine all the named entity chunks to one whole chunk. The result is as below.

```
Named entity recognition and combining chunks:
['O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-per', 'O', 'O', 'O', 'O', 'O', 'O', 'B-geo', 'O']

[['I', 'O'], ['will', 'O'], ['meet', 'O'], ['my', 'O'], ['sister', 'O'], [',', 'O'], ['Alice', 'per'], [',', 'O'], ['at', 'O'], ['3', 'O'], ['pm', 'O'], ['maybe', 'O'], ['around', 'O'], ['Motosu', 'geo'], ['.', 'O']]

```

Picture III.3.7 Example 2 step 2

The third step is to detect the private locational candidate phrases and do anonymization with another subdivision/state that belongs to the same country.

```
Private locational candidate_phrases:
[['i', u'will', u'meet', 'my', u'sister', ',', 'Alice', ',', 'at', '3', 'pm', u'mayb', u'around', 'Motosu']]

Anonymize private locational phrases:
[['i', '0'], [u'will', '0'], [u'meet', '0'], ['my', '0'], [u'sister', '0'], [',', '0'], ['Alice', 'per'], [',', '0'], ['at', '0'], ['3', '0'], ['pm', '0'], [u'mayb', '0'], [u'around', '0'], ['Takayama', 'geo'], ['.', '0']]
```

Picture III.3.8 Example 2 step 3

The fourth step is to detect the private personal candidate phrases and do the anonymization. The result is as below.

```
Private personal candidate_phrases:
[['i', u'will', u'meet', 'my', u'sister', ',', 'Alice'], ['my', u'sister', ',', 'Alice']]

Anonymize private personal phrases:
[['i', '0'], [u'will', '0'], [u'meet', '0'], ['my', '0'], [u'sister', '0'], [',', '0'], ['Anna', 'per'], [',', '0'], ['at', '0'], ['3', '0'], ['pm', '0'], [u'mayb', '0'], [u'around', '0'], ['Takayama', 'geo'], ['.', '0']]
```

Picture III.3.9 Example 2 step 4

The fifth step is to convert list of normalized token back into a text. We have to also convert back all the words that had been converted to their lemma back to the original words. The last step is to compute sentence similarity to calculate the information loss. The result is 0,69 which means the two sentences have a quite high similarity and a medium information loss. The result is as below.

```
Private temporal phrases :
[['at', '3', 'pm']]

Final anonymized sentence:
I will meet my sister , Anna , in the afternoon maybe around Takayana .

Similarity measure value:
0.697738095238
nim_13514065@Gerald:~/Desktop/GNCT Work/CRFAnonymization$
```

Picture III.3.10 Example 2 step 5

There are still many cases when the rule based approach fails. For example, if there is any private verbs/nouns that are currently not listed in the database, then it will not be classified as private verbs/nouns by the rule. This case sometimes can be covered by the co-occurrence metrics step if the noun has high co-occurrence metrics value with the user profile but sometimes can't if the condition is vice versa (low co-occurrence metrics value).

Another case is when rule based approach fails is when there is a word that refers to another word (e.g. he, she, it). Currently, the rule based approach ignore all the phrases starting with he/she/it because I haven't got time to work on co-reference resolution to get information about the word that is referred by he/she/it. An example of text where the rule based approach fails is described in paragraph below.

.

The text : “My sister’s name is Atika. She currently lives in London for higher education.” In this case, London will not be anonymized because it is not detected as candidate phrase (not between my/I and location). But, it is wrong since London is a private information of the user’s sister.

Chapter IV

Closing

This chapter contains two closing parts, which are the conclusion and suggestion. The conclusion is derived from the research project the student has completed and also from the work environment in NIT-GC. Meanwhile, the suggestion is about some future works and/or improvements that can be accomplished and the student's feedback regarding the industrial practice.

IV.1 Conclusion

To sum up, I want to give two kind of conclusions, one is about the research substance and the other one is about the industrial practice at National Institute of Technology, Gifu College (NIT-GC). Below is for the research substance.

1. To create good named entity recognizer, corpus dataset plays very important roles. Make sure you choose a corpus that suits your problem domain (e.g. in my case I need to make sure that my corpus have sufficient numbers of person, location, and organization entities).
2. Another important factor is the algorithm and the feature you use to train your classifier. In my experiments, conditional random field algorithm did better than other algorithms because it concerns not only a word that is going to be predicted, but also concerns the word before and the word after (linear chain CRF).
3. The private phrases detection still have so many flaws to cover. This is because there are many other patterns that are not included/covered in the rule-based approach method. Another reason is a limited private

and non-private verbs/nouns in the database. The private and non-private verbs/nouns corpus should be extended to achieve better performance.

Meanwhile, below is for the work environment at NIT-GC.

1. Overall, the industrial practice at NIT-GC is both challenging and fun.
2. It's challenging because the supervisor always gives you feedbacks and urges you to improve your work.
3. It's fun of course because it's Japan. But, it's more because all the students in the lab are very nice and helpful. They invited me to play board games and to eat some delicious Japanese dishes at the restaurants. They also took me to some beautiful cities in Japan, such as Kyoto, Osaka, and Nagoya in our free time (weekend).
4. It's fun also because they make the lab as a very nice workplace. Beside computer facilities, they have many entertainment stuffs such as plenty of board games to play. Even the lecturer play the board games and he just blend with his students as if he is their friends.

IV.2 Suggestion

I want to give two kind of suggestions, one is about the research substance and the other one is about the industrial practice at National Institute of Technology, Gifu College (NIT-GC). Below is for the research substance.

1. There are currently not many anonymization research focusing on raw text. Thus, I hope that this research could be extended furthermore in the future in term of improving performance by applying other methods and also in term of domains/areas. I hope that I can also create similar

researchs for other domains, such as for raw company or government documents.

2. One of the solution that may improves the private phrases detection is to apply co-reference resolution to the private named entity detection step. It basically will determine all the words that refer to the same thing and that may improves the private named entity detection performance. Other solutions are of course to define more rules in the rule-based method and to extend the private and non-private verbs/nouns corpus.

3. One of the solution that may improves the named entity recognition (especially for event entity) is to redefine the feature used in the CRF classifier and to supply more annotated sentences from other dataset (not just GMB dataset).

Meanwhile, below is for the industrial practice at NIT-GC.

1. For ITB exchange students who are going to do internship next year, it may be good if they are invited to attend some relevant lectures/classes and are involved in student's experiment activities.

2. It may be good for next year if NIT-GC opens internship opportunity for other departments like Mechanical Engineering, Civil Engineering, or Architecture so it can cover more ITB exchange students.

References

- [1] Nguyen Son Hoang Quoc. Anonymizing Private Phrases and Detecting Disclosure in Online Social Networks, 2015.
- [2] Rada Mihalcea and Courtney Corley. Corpus-based and Knowledge-based Measures of Text Semantic Similarity, 2006.
- [3] Latanya Sweeney. k-Anonymity: A Model for Protecting Privacy, 2002.
- [4] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python, 2009.
- [5] Aurelien Geron. Hands-On Machine Learning with Scikit-Learn and Tensorflow, 2017.